The OryzaSNP Project – Genome-wide SNPs discovered in diverse Rice

Kenneth L. McNally[1], Kevin L. Childs[2,3], Regina Bohnert[4], Keyan Zhao[5], Victor Ulat[1], Richard Clark[6], Georg Zeller[4,6], Douglas Hoen[7], Thomas Bureau[7], Renee Stokowski[8], Dennis Ballinger[8], Kelly Frazer[8], David Cox[8], Badri Padhukasahasram[5], Carlos D. Bustamante[5], Gunnar Rätsch[4], Detlef Weigel[6], Richard Bruskiewich[1], David Mackill[1], C. Robin Buell[2,3], Rebecca Davidson[9], Jan Leach[9], and Hei Leung[1]

[1]International Rice Research Institute, DAPO Box 7777, Metro Manila 1301, the Philippines;
[2]The Institute for Genomic Research, 9712 Medical Center Dr., Rockville, MD 20850, U.S.A.;
[3]Department of Plant Biology, 166 Plant Biology Building, East Lansing MI 48824, U.S.A.;
[4]Friedrich Miescher Laboratory of the Max Planck Society, D-72076 Tübingen, Germany;
[5]Department. of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853 U.S.A.;
[6]Department of Molecular Biology, Max Planck Institute for Developmental Biology, D-72076 Tübingen, Germany
[7]Department of Biology, McGill University, 1205 Dr. Penfield Av., Montreal, Quebec H3A 1B1, Canada;
[8]Perlegen Sciences, Inc., 2021 Stierlin Court, Mountain View, CA 94043, U.S.A.;
[9]Bioagricultural Sciences and Pest Management, Colorado State University, Ft. Collins, CO 80523, U.S.A.

Abstract

The OryzaSNP project will undertake genome-wide SNP discovery across a 100 Mb fraction of the Nipponbare rice genome with little or no repetitiveness for 20 diverse varieties. The diverse varieties included representatives from all variety groups with Nipponbare included as control. SNPs are being identified by array-based re-sequencing technology using very high-density oligomer arrays at Perlegen Sciences in two phases: a development phase and a discovery phase. In the development phase, a small genomic region located on Chr. 3 from 28.2 to 28.9 Mb was used to optimize experimental conditions. Model-based (MB) algorithms predicted a set of 2123 redundant (nr) SNPs in the development region, sites in Nipponbare where the nucleotide differed in one or more of the other 19 varieties. In the development region, 3.2 SNPs per kb were found, consistent with figures obtained from pair-wise comparisons of indica and japonica varieties. On the genome-wide dataset, we will be applying the machine learning (ML) algorithms developed for the *Arabidopsis* SNP discovery project to predict another set of SNPs. Annotation relative to the Rice Annotation Project release 2 and TIGR release 5 gene models on the full dataset will be available via the OryzaSNP consortium website. Additional analyses for understanding the extent of LD and genes/regions introgressed from one type to another or potentially involved in domestication will be undertaken. As a second phase of the OryzaSNP project, a consortium of partners is planned to design high-density genotyping arrays containing the discovered SNPs and other features for genotyping a large collection of over 2000 varieties.

Media Summary

The OryzaSNP consortium has undertaken genome-wide SNP discovery on a diverse collection of 20 rice varieties by high-density oligomer array-based re-sequencing.

Keywords

Oryza sativa, Single Nucleotide Polymorphism, linkage disequilibrium

Introduction

Rice is the world's most important cereal crop feeding half of humanity. With the publication of the high-quality genomic sequence for the temperate japonica variety Nipponbare by the International Rice Genome Sequencing Project in 2005, an unprecedented opportunity to discover genome-wide single nucleotide polymorphisms (SNPs) was at hand. Such a SNP collection will allow better understanding of the extent of linkage disequilibrium and haplotype structure in an inbred crop, identification of SNPs for genes or regions underlying important traits for crop improvement through targeted or genome-wide association studies, and supply a large collection of new markers for plant breeding. The re-sequencing technology pioneered by Perlegen Sciences, Inc. is an array-based re-sequencing method where variation is detected by DNA-DNA hybridization using very high-density oligomer arrays. They have successfully applied it for SNP discovery in the human, mouse and *Arabidopsis* genomes (Hinds et al, 2006; Frazer et al 2007; Clark et al 2007). The OryzaSNP project has involved two phases, the first phase involved SNP discovery on a short region of Chr.

3 while the second phase spanned the genome. Here, we report preliminary analyses on SNPs discovered in the development phase region by model-based algorithms.

Methods

Twenty diverse rice varieties spanning the range of varietal groups with representatives of traditional, advanced and improved lines were chosen for re-sequencing (Table 1). These varieties underwent one round of purification from a single plant. Resulting seed was grown for DNA extraction and further increase.

Table 1. Diverse rice varieties re-sequenced by high density array-based technology.

IRIS (Code) [a]	Variety	Source	Origin	Group [b]
2254728 (2)	Nipponbare	IRTP[c] 23787	Japan	Temp[d] japonica
2021623 (3)	Tainung 67	Academia Sinica	Taiwan	Temp japonica
2254732 (7)	Li-Jiang-Xin-Tuan-Hei-Gu (LTH)	IRGC[e] 59323	China	Temp japonica
2254738 (4)	M 202	IRGC 77142	U.S.A.	Temp japonica
2254730 (5)	Azucena	IRTP 4209	Philippines	Trop[f] japonica
2254722 (6)	Moroberekan	IRGC 12048	Guinea	Trop japonica
2254737 (9)	Cypress	IRRI PBGB[g]	U.S.A.	Trop japonica
2254723 (8)	Dom-sufid	IRGC 12880	Iran	Aromatic[h]
2254720 (11)	N 22	IRGC 4819	India	Aus
2254725 (14)	Dular	IRGC 32561	India	Aus
2254721 (13)	FR13 A	IRGC 6144	India	Aus
2254726 (12)	Rayada	IRGC 77210	Bangladesh	Deep-water 4
2254724 (10)	Aswina	IRGC 26289	Bangladesh	Deep-water 3
2254729 (21)	IR64 (IR64-21)	IRRI PBGB	Philippines	Indica
2254731 (18)	Shan-Huang Zhan-2 (SHZ2)	IRTP 16338	China	Indica
2254727 (19)	Pokkali	IRGC 108921	India	Indica
2254736 (20)	Swarna	IRRI PBGB	India	Indica
2254719 (17)	Sadu-cho	IRGC 2243	Korea	Indica
2030504 (15)	Minghui 63	Huazhong Ag. Un.	China	Indica
2030525 (16)	Zhenshan 97B	Huazhong Ag. Un.	China	Indica

[a]The germplasm identifier (GID) in the International Rice Information System (IRIS, iris.irri.org) and numerical code assigned for analyses; [b]Determined by isozyme, SSR, SNP analyses, or a combination thereof; [c]International Network for the Genetic Evaluation of Rice accession; [d]Temperate type; [e]International Rice Genebank Collection accession; [f]Tropical type; [g]Plant Breeding, Genetics and Biotechnology (Division); [h]basmati (aromatic) plant type.

For the development phase, a region from 28.2 to 28.9 Mb on Chromosome 3 from the IRGSP v4 pseudomolecules was used (IRGSP, 2005). This region was masked for repetitive regions using the TIGR repeat database (Yuan et al 2003) and the Rice Transposable Element database (Juretic et al 2004). Long-range PCR primers (76 pairs) were designed to amplify fragments across the entire region. A high density oligomer array was produced by Affymetrix by the tiling approaches previously described (Patil et al, 2001; Frazer et al, 2007; Clark et al, 2007). LR-PCR amplicons (median size 9.4 kb) were produced for the 20 varieties, pooled into fractions, and sheared. Sheared fragments were fluorescently labeled and hybridized in pools of two varieties per array, similar to what used for mouse SNP discovery (Frazer et al, 2007). A set of 2123 SNPs was predicted from hybridization patterns using model-based algorithms (Frazer et al, 2007).

Results

Conditions for hybridization rice LR-PCR amplicons were optimized using the development array with a portion of Chr. 3 located from 28.2 – 28.9 Mb (IRGSP r4 coordinates). In this region, 2123 nr MB SNPs were identified. The population structure of the 20 varieties plus the Nipponbare reference sequence was estimated by model-based inference using Structure 2.2 (Pritchard et al, 2000). The best K-value was 2 by the Evanno Criterion of maximal ΔK (Evanno et al, 2005) corresponding to a split between japonica/ aromatic and aus/indica/deep-water types. However, the tropical japonica variety, Cypress, placed in the indica group. The model-based population structure for inbreeding species, InStruct, was also applied (Gao et al 2007) with K=3 having the minimal Deviance Information Criterion (Figure 1). For the groupings from InStruct (K=3), Cypress grouped with the indicas and Minghui 63 (indica type) grouped with the aus. For the model-based inferences, five chains were run for each K with a burnin of 50,000 followed by 150,000 iterations for sampling; further, Structure was run on phased haplotype data with the admixture model and InStruct was run using diploid data assuming 3 populations and admixture. The population structure from InStruct at K=3 was supported by clustering obtained from both DNA parsimony and Neighbor Joining using Phylip 3.67 (Figure 2). The SNPs used for these analyses occur in 700 kb window; hence, the groupings obtained from phylogenetic or population genetic analyses, represent a snapshot of the history of these 20

varieties. Therefore, placement in a group other than the expected one is probably due to their breeding histories: the pedigree of Cypress includes IR 8 (indica) as a parent while Minghui 63 contains "GuyanasDao" of an unknown variety type.

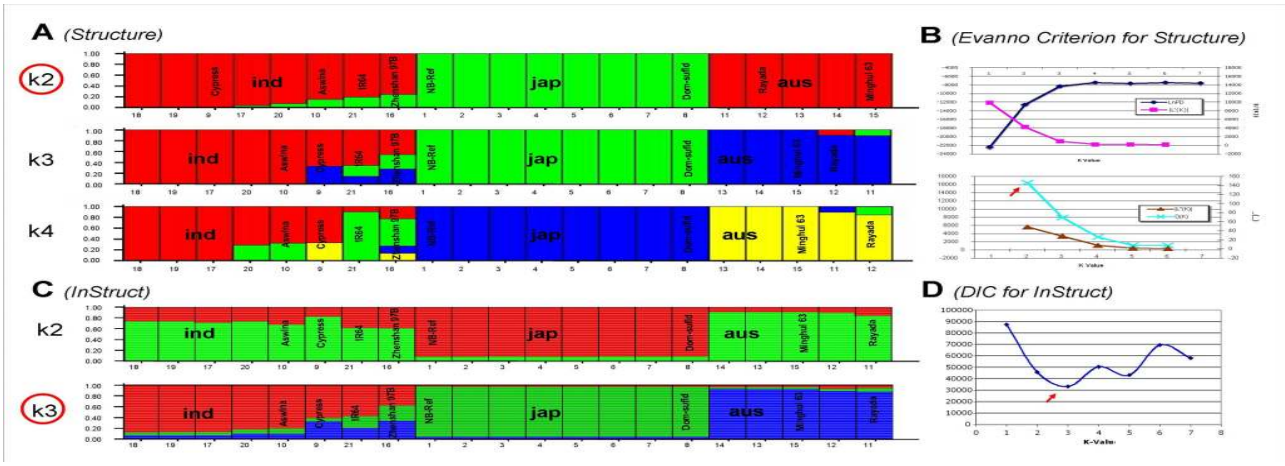

Figure 1. Groupings by Structure (A)and InStruct (C) analyses for the region from 28.2 – 28.9 Mb on Chr. 3. The best k-value by the Evanno criterion for Structure was k=2 (maximum ΔK) while the best k value for InStruct was k=3 (minimal DIC value) (red circles and arrows).

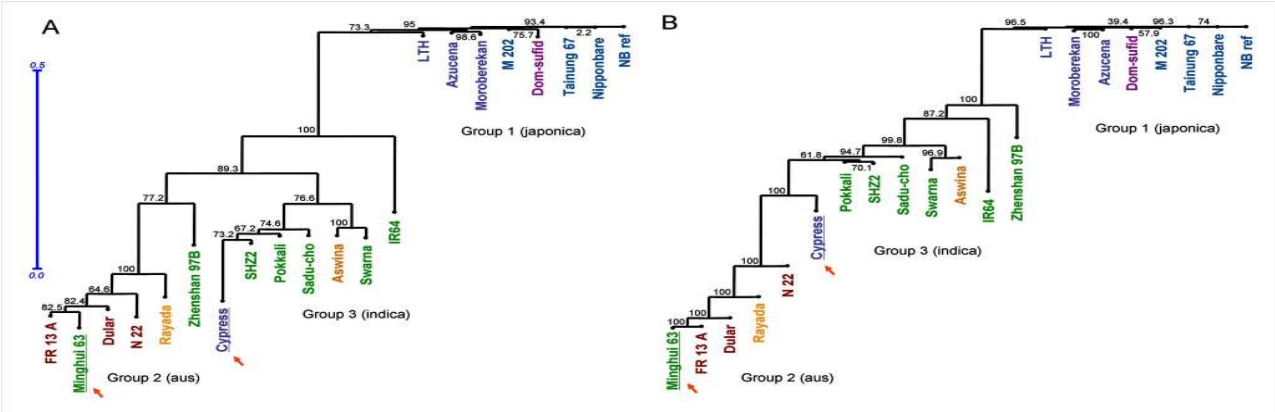

Figure 2. Trees by DNA parsimony (A) and Neighbor Joining (B) by Phylip 3.67 with 1000 bootstrap resamplings and 10 jumblings per bootstrap sample. Cypress (tropical japonica) groups with indica and Minghui 63 (indica) groups with aus (red arrows).

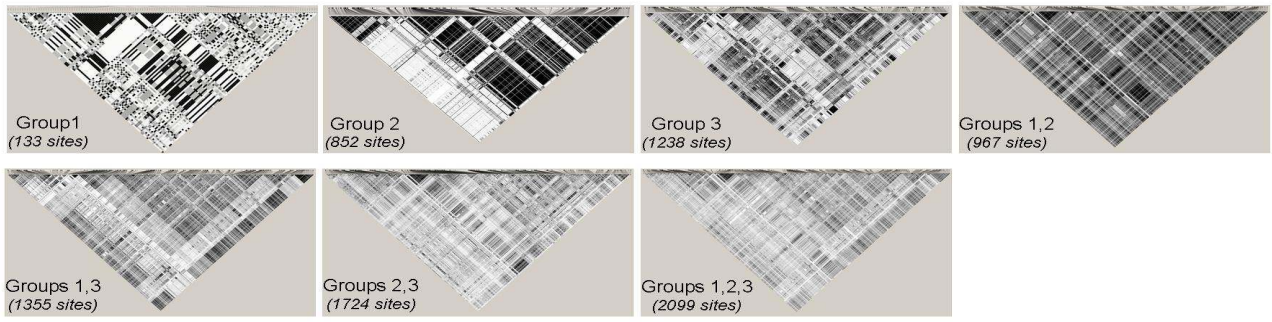

Figure 3. LD (r^2) was estimated in the region from 23.8 to 30.8 Mb on Chr. 3 for within and between group comparisons using HaploView 3.32. Apparent LD in this 700 kb window extends to 300 kb or longer. White represents an r^2 of 0 (linkage equilibrium) while black corresponds to an r^2 value of 1 (linkage disequilibrium).

In an initial attempt to understand LD, we calculated pairwise r^2 values for SNPs by Haploview 3.32 within and between each of the three groups (Figure 3). LD in this region of Chr. 3 appears to extend for at least 300 kb or longer at the 3'-end, with the strongest LD observed between the japonica and aus groups.

Conclusion

For the SNPs discovered during the development phase, we identified numerous SNPs in the rice genome relative to Nipponbare, indicating that the discovery of SNPs in the 100 MB fraction of the genome will provide a foundation on which other studies can now be built. We plan other analyses to define the number of blocks having shared SNPs from one of the three groups in the background of another group. Shared regions should be indicative of a common breeding history or, possibly, they might be related to the domestication of a particular group or type relative to the others. We are also in the process of analyzing the possible effect of SNPs on gene and protein function. OryzaSNP2, a project to undertake high-density genotyping on 2000+ varieties has been initiated with expressed interest from consortium partners (IRRI, Cornell, USDA, Korea, India, CIRAD, Academia Sinica, Max Plank Institute, and, potentially, additional partners). High-density Affymetrix genotyping arrays using OryzaSNP data and SNP data from other projects will be designed. The resulting genotype information when combined with detailed phenotype data will allow the global rice community to undertake genome scans and perform association studies. These studies have the prospect of delivering to plant breeders useful alleles at genes underlying traits of importance for plant improvement, e.g. those involved in the expression or regulation of heretofore intractable traits such as tolerance to drought.

References

Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, et al. 2007. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science 317:338-342

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software Structure: A simulation study. Mol. Ecol. 14:2611–2620.

Felsenstein, J. 2004. PHYLIP (Phylogeny Inference Package) version 3.67. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, Gupta RV, Montgomery J, Morenzoni MM, Nilsen GB, et al. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature 448:1050-1053

Gao H, Williamson S, Bustamante CD. 2007. A Markov Chain Monte Carlo approach for the joint inference of population structure and inbreeding rates form multilocus genotype data. Genetics 176:1635-1651

Haploview 3.32. 2006. http://www.broad.mit.edu/mpg/haploview

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307:1072-1079

Juretic N, Bureau TE, Bruskiewich RM. 2004. Trasnposable element annotation of the rice genome. Bioinformatics 20:155-60.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. Nature 436**:**793-800.

International Rice Information System <http://iris.irri.org>

OryzaSNP. 2007. Site hosted at the International Functional Genomics Consortium <http://irfgc.irri.org>

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, et al. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719-1723

Pritchard JK, Stephens M, Donnelly. 2000. Inference of population structure from multilocus genotype data. Genetics 155:945-959.

Yuan Q, Ouyang S, Liu J, Suh B, Cheung F, Sultana R, Lee D, Quackenbush J, Buell CR. 2003. Nucleic Acids Res 31:229-233.